Prediction of PIK3CA mutation with gene expression

Jun Kang [*]

[*] Corresponding author: jkang.alien@gmail.com

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur eget porta erat. Morbi consectetur est vel gravida pretium. Suspendisse ut dui eu ante cursus gravida non sed sem. Nullam sapien tellus, commodo id velit id, eleifend volutpat quam. Phasellus mauris velit, dapibus finibus elementum vel, pulvinar non tellus. Nunc pellentesque pretium diam, quis maximus dolor faucibus id. Nunc convallis sodales ante, ut ullamcorper est egestas vitae. Nam sit amet enim ultrices, ultrices elit pulvinar, volutpat risus.

Introduction

A list

- Brevity
- Logic and clarity
- Clean typing

The problem

- PIK3CA mutation in selecting drug

Breast cancer with PIK3CA mutation has been approved to use PIK3CA inhibitor in hormone receptor positive HER2 negative subtype. [1] Prediction of PIK3CA mutation was done by gene expression data of TCGA.

- Predicting mutation by gene expression data
  - What for?
  - Feasible?
  - Cost?
- Varing frquency of PIK3CA mutation across cancer types
  - High in endometrial breast
  - not common other cancer type

Universial prediction using gene expression data across cancer type for certain kind of mutation. It’s clinically useless now, but we want to explore the possibilities of the PIK3CA mutation prediction.

RNAseq can be widely used. The mutation status directly prediction

Previous study [2]

prediction of aberant activation of a certain pathway vs a specific mutation The proposed solution

Prediction RAS pathway activation with gene expression data was done in previous study. [2] They trained pancancer The Cancer Genome Atlas (TCGA) data with a supervised elastic net penalized logistic regression classifier with stochastic gradient descent. The performance of their model was 84% with an area under the receiver operating characteristic (AUROC) curve and 63% with an area under the precision recall (AUPR) curve. We applied their modeling methods in prediction PIK3CA mutation.

Materials and Methods

Dataset

We used TCGA pancancer dataset. PIK3CA mutation data was get using cgdsr rpackage. Gene expression data was get from GDAC firehose using RTCGAToolbox R package. ER immunostain postive HER2 immunostain negative and/or SISH negative breast cancers are included. Data of invasive ductal carcinomas were used for training set and data of invasive lobular carcinoma were used for test set. Number of observations were 530 in training set and 188 in test set.

Selecting variable for modeling

To narrow down potential predictors, Genes with a large the median absolute deviation (more than third-quartiles) were selected. 5000 out of 20502 genes were included in the modeling process.

Preprocessing

Yeo-Johnson transformation was done to correct skewness. Centering and scaling were done. All preprocessing was done using recipe r package.

Modeling

Penalized logistic regression was applied to prediction modeling. 10-fold cross-validation with targe variable stratification was done over the hyperparameter grid: λ $\{10^{-5}$, 10^{-4},10^{-3},10^{-2},10^{-1}, $10^{0}\}$, α $\{0.0, 0.25, 0.5, 0.75\}$. Model performance was evealuated with the area under receiver operating characteristic (ROC) and area under the ROC (AUROC) and precision recall (PR) and area under curve (AUPR).

Results

selecting model

Cross-validation

The model showed best performance at lambda = 0.01 and alpha = 1.0 (Ridge regression).

Prediction performance

performance of each cancer type

Important predictors

Discussion

Our model showed good performance to predict PIK3CA mutation of various cancer types.

Main message answers the question and main supporting evidence

Critical assessment opinions on - any shortcomings in study design

In this study, the split method was done for testset instead of indepedent testset. Because there is no standard analysis method for the RNA-seq gene expression quantification, it is difficult to find gene expression data same as TGCA data. Developing standard analysis method for gene expression data is neccessory to apply this prediction model.

- limitations in methods

Case imbalance

- flaws in analysis
- validity of assumption

Comparison with other studies where inconsistencies are discussed

Conclusions comments on possible biological or clinical implications and suggestions for further research.

Evaluate the results - not the authors

References

Figure legends

1. André F, Ciruelos E, Rubovszky G, Campone M, Loibl S, Rugo HS, et al. Alpelisib for PIK3CA-Mutated, Hormone ReceptorPositive Advanced Breast Cancer. New England Journal of Medicine. Massachusetts Medical Society; 2019;380: 1929–1940. doi:10.1056/NEJMoa1813904

2. Way GP, Sanchez-Vega F, La K, Armenia J, Chatila WK, Luna A, et al. Machine Learning Detects Pan-cancer Ras Pathway Activation in The Cancer Genome Atlas. Cell Reports. Elsevier; 2018;23: 172–180.e3. doi:10.1016/j.celrep.2018.03.046